ICDAR2019 Competition on Historical Book Analysis -HBA2019

Maroua Mehri, Pierre Héroux, Rémy Mullot, Jean-Philippe Moreux, Bertrand B. Couasnon, Bill Barrett

ICDAR2019 Competition on Historical Book Analysis - HBA2019

Maroua Mehri*†, Pierre Héroux†, Rémy Mullot‡, Jean-Philippe Moreux§, Bertrand Coüasnon¶ and Bill Barrett‖
*Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023 Sousse, Tunisie
†Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
‡University of La Rochelle, L3i Laboratory, 17042 La Rochelle, France
§Bibliothèque nationale de France, Digitization service, 75706 Paris, France
¶Univ Rennes, CNRS, IRISA, F-35000 Rennes, France
‖Brigham Young University, 84602 Provo, USA
Emails: maroua.mehri@eniso.u-sousse.tn, pierre.heroux@univ-rouen.fr, remy.mullot@univ-lr.fr,
jean-philippe.moreux@bnf.fr, bertrand.couasnon@irisa.fr and barrett@cs.byu.edu

***Abstract*—In this paper, we present an evaluative study of pixel-labeling methods using the HBA 1.0 dataset for historical book analysis. This study is held in the context of the 2nd historical book analysis (HBA2019) competition and in conjunction with the 15th IAPR international conference on document analysis and recognition (ICDAR2019). The HBA2019 competition provides a large experimental corpus and a thorough evaluation protocol to ensure an objective performance benchmarking of pixel-labeling document image methods. Two nested challenges are evaluated in the HBA2019 competition: *Challenge 1* and *Challenge 2*. *Challenge 1* evaluates how image analysis methods could discriminate the textual content from the graphical ones at pixel level. *Challenge 2* assesses the capabilities of pixel-labeling methods to separate the textual content according to different text fonts (e.g. lowercase, uppercase, italic, *etc.*) at pixel level. During the competition, we received 52 and 38 different teams' registrations for *Challenge 1* and *Challenge 2*, respectively and finally 5 of them submitted their results in each challenge. Qualitative and numerical results of the participating methods in both challenges are reported and discussed in this paper in order to provide a baseline for future evaluation studies in historical document image analysis. The evaluation shows that the method submitted by the NLPR-CASIA team achieves the highest performance in both challenges.**

***Keywords*-HBA 1.0 dataset, Historical book analysis, Pixel-labeling.**

I. Introduction

Over the last few years, libraries and archives have conducted large digitization programs with cultural heritage documents to guarantee lasting preservation and worldwide access to historical collections. The increasing interest in digital libraries of Google and recently of other leaders and large firms (e.g. Microsoft, IBM, Yahoo and Amazon) proves the major success and effervescence of digital libraries and their rapid growth worldwide, and it poses new specific challenges concerning the preservation and reproduction of historical collections to reinforce its leadership position [1]. Providing robust and accurate historical document image analysis (HDIA) methods has been identified as among the major challenges [2]. Indeed, HDIA remains an open issue due to the idiosyncrasies of historical documents, such as the superimposition of information layers (e.g. stamps, handwritten notes, noise, back-to-front interference and page skew) and the variability of their contents and/or layouts, *etc.* Moreover, HDIA will often be complicated by the unavailability of *a priori* knowledge about the layout and content, scanning resolution or DI size, *etc.* On the other side, factors concerning the book edition and digitization properties (e.g. paper quality, printing defects, degradation, black borders, darker areas in the binding margins, flat scan by opening pair of pages, noise generated by the scanner roller and sensor, contrast/brightness level, curvature and compression) further complicates the HDIA task. Therefore, the research community specialized in HDIA is continuing to investigate and provide efficient historical document image segmentation, layout analysis and characterization methods.

Since 2011, many competitions have been proposed for historical document image segmentation and layout analysis in the context the ICDAR and the ICFHR conferences [2]. These competitions are hindered by many issues related to the different provided datasets of historical document images. First of all, the different datasets of historical document images provided in the context of theses contests focus on either a specific kind of document such as historical newspaper layout analysis (HNLA), or a specific application such as handwritten text recognition (HTRtS and RHHT), multi-spectral text extraction (MS-TEx), text line detection (ANDAR-TL), word recognition (ANWRESH), keyword spotting (KWS), classification of medieval handwritings in Latin script. Moreover, the majority of these datasets are composed of pages having similar content and layout characteristics or collected from a single book or collection such as the GERMANA corpus or the RODRIGO corpus [3]. On the other side, there is a limited number of realistic, comprehensive and flexibly structured datasets of historical document images and their associated ground truths. In addition to these issues which are critical to provide an informative benchmarking of HDIA methods, the lack of a common dataset of historical document images and

the appropriate quantitative evaluation measures are also highlighted by researchers.

Therefore, we present in this paper the report of the 2nd edition of the historical book analysis (HBA2019) competition[1] which is sponsored by the 15th IAPR international conference on document analysis and recognition (ICDAR2019). The HBA2019 competition provides a comparative study of five pixel-labeling methods using a large experimental corpus of historical document images, the HBA 1.0 dataset [4] and an evaluation protocol to address specific issues related to HDIA methods for a collection of several digitized historical books. The HBA2019 competition addresses an important lack in past competitions by providing a means of consistent evaluation and comparison of pixel-based image analysis methods for digitized historical book analysis. The goal of the HBA2019 competition consists of showing that low-level image analysis methods can be tuned on a small training dataset (*i.e.* a small number of book pages with their associated ground truth provided) in order to deduce automatically the corresponding information on the remaining pages of the analyzed book. One key feature of the proposed experimental corpus for the HBA2019 competition is that it is composed of images that represent all pages of books. Moreover, the HBA2019 competition also evaluates the generalization and adaptation capabilities of low-level image analysis methods as they are applied on a dataset that presents a large variety of ancient books.

The remainder of this paper is organized as follows. Section II presents an overview of the HBA2019 competition and its two challenges. In Section III, a brief description of the HBA 1.0 dataset is introduced. Section IV details the evaluation protocol by outlining the ground truth, the experimental protocol and the accuracy metrics used to evaluate the performance of each participating method in the HBA2019 competition. Each participating method in both challenges of the HBA2019 competition is summarized in Section V. Qualitative and numerical results of the evaluation of the participating methods are reported and discussed in Section VI. Finally, our conclusions and future work are presented in Section VII.

II. HBA2019 COMPETITION

The HBA2019 competition is focused on evaluating the discriminating power of pixel-labeling methods for segmenting graphical contents from textual ones on the one hand, and for separating various text fonts and scales on the other hand, in images representing the pages of historical books. Therefore, two nested challenges are proposed in the HBA2019 competition. The first challenge (*Challenge 1*) is interested in raising issues related only to how image analysis methods perform for discriminating the textual content from the graphical ones. In the *Challenge 1*, a binary classification task is evaluated. On the other side, the second challenge (*Challenge 2*) evaluates the capabilities of pixel-labeling methods to firstly distinguish between text and graphic, and secondly to separate the textual content according to different text fonts (e.g. lowercase, uppercase and italic). In the *Challenge 2*, a multi-class classification task is assessed. Indeed, the textual content can contain formatting such as many different typefaces and sizes which are associated to the structure level of the analyzed document. The logical level aims to interpret and recognize the different parts that compose a document image and specify the logical relationship between them (e.g. body text, legend, annotation and chapter title).

III. HBA 1.0 DATASET

To provide an informative benchmarking of HDIA methods, many researchers have addressed the need of a realistic (*i.e.* it must be composed of real digitized document images), comprehensive (*i.e.* it must be well characterized and detailed for ensuring in-depth evaluation) and flexibly structured (*i.e.* to facilitate a selection of sub-sets with specific conditions) dataset. Nevertheless, representative datasets of historical document images with their associated ground truths are currently hardly publicly accessible for HDIA. This is mainly due to the intellectual and industrial property rights on the one hand, and the definition of an objective and complete ground truth of a representative dataset of historical document images on the other hand [2]. Thus, in this study we propose to use the dataset associated with the HBA2019 competition[1], which is the HBA 1.0 dataset. Figure 1 illustrates few book page samples of the HBA 1.0 dataset. More details can be found in [4].

(a) ***Book 02*** (b) ***Book 05*** (c) ***Book 06*** (d) ***Book 07*** (e) ***Book 10***

Figure 1. Sample book pages of the HBA 1.0 dataset.

The HBA 1.0 dataset which was introduced in [4], is composed of 4,436 real scanned ground truthed one-page historical document images from 11 books (5 manuscripts and 6 printed books) in different languages and scripts published between the 13th and 19th centuries. The images composing the HBA 1.0 dataset have been collected from the French digital library Gallica[2]. The HBA 1.0 dataset remains exclusive property of the BnF. It is hosted on a server maintained by the HBA2019 competition organizers and will be made publicly available (along with the ground truth) for scientific research use. It will be among the first

[1] http://hba.litislab.eu/

[2] http://gallica.bnf.fr

of publicly released datasets for layout analysis containing annotations at pixel level.

IV. Evaluation protocol

In this section, we briefly present the evaluation protocol used to assess the performances of the participating methods in both challenges of the HBA2019 competition by outlining the defined ground truth, the experimental protocol and the used accuracy metrics. More details can be found in the HBA2019 competition Website[1].

A. Ground truth

A set of classes has been used for each book of the HBA 1.0 dataset to define its ground truth. The ground truth of the HBA 1.0 dataset is obtained by annotating each foreground pixel. The ground truth of each selected foreground pixel has been defined by means of a label indicating the content type/class of the analyzed historical document image. Different labels for the selected foreground pixels with different fonts have been also assigned for evaluating pixel-labeling methods to separate various text fonts. Figure 2 illustrates few examples of ground truthed book pages of the HBA 1.0 dataset. Each selected foreground pixel is marked by a color that symbolizes the corresponding content type. The ground truth of the HBA 1.0 dataset contains more than $7,58$ billion annotated pixels.

(a) ***Book 01*** (b) ***Book 03*** (c) ***Book 06*** (d) ***Book 07*** (e) ***Book 10***

Figure 2. Sample ground truthed book pages of the HBA 1.0 dataset.

B. Experimental protocol

The HBA 1.0 dataset is divided into two sub-datasets, the sample and evaluation datasets. The sample dataset contains two books while the evaluation one is composed of the nine remaining books. Each dataset is composed of a set of training images and a set of test images. The training dataset contains a reduced number of book pages, along with their ground truth. The training images are representative of different contents and layouts of the book pages. On the other side, the test dataset is composed of images representing the remaining book pages. The sample dataset is provided in order to fine-tune the participating methods in the HBA2019 competition (*i.e.* training or testing).

All pages of the two selected books composing the sample dataset, along with their ground truths are provided. A few pages are selected from each book of the sample dataset to constitute the training dataset. These selected pages should contain all content classes of the analyzed books. The two selected books that form the sample dataset are: *Book 04* as a manuscript book and *Book 06* as a printed one. The nine selected books that form the evaluation dataset are: *Book 01*, *Book 02*, *Book 03* and *Book 05* as manuscript books, and *Book 07*, *Book 08*, *Book 09*, *Book 10* and *Book 11* as printed ones. Only the ground truths of the training images of the nine selected books of the evaluation dataset are provided.

The HBA2019 competition aims at evaluating methods, which would automatically annotate an important number of book pages, based on a limited number of manually annotated pages of the same book. The provided limited number of manually annotated pages constitutes the training image dataset. Therefore, it is ensured that each class of content type is represented in the set of the training pages for each book. It is worth pointing out that the content classes in the HBA 1.0 dataset vary from one book to another book and have very different headcounts. Indeed, the textual content is predominant in monographs, compared with the graphical content. Moreover, among the textual content a great majority represent the body text while other character fonts are more marginal. This imbalanced headcounts between classes varies from one book to another book in the HBA 1.0 dataset. There is surely a great deal of imbalanced headcounts between classes in the same book. The class headcounts in the training dataset are not thereby similar to the test dataset. Indeed, the minority classes are adequately represented in the training dataset in order to ensure an appropriate learning task. However, unlike the minority classes, the majority classes are clearly less represented in the training dataset in comparison with the class headcounts in the test dataset. These requirements have been satisfied in the selection of the training pages in the sample dataset. Similarly, these rules are also applied to the selection of the training pages in the evaluation dataset. It is worth noting that the prime difficulty of the experimental protocol of the HBA2019 competition lies in the experimental corpus which contains different types of content in historical books published at different eras such as printed books from the 19^{th} century or manuscripts from the 13^{th} century.

C. Evaluation metrics

To evaluate the performance of the participating methods in the HBA2019 competition, the confusion matrix is computed based on the ground truth and the obtained results. From the confusion matrix, several per-pixel classification accuracy metrics are computed [5].

- The classification accuracy rate (CA) metric corresponds to the ratio of the true classified predicted pixels and the total number of pixels.
- The F-measure (F) corresponds to a score resulting from the combination of the precision and recall accuracies by using a harmonic mean. It assesses both the homogeneity and completeness criteria of a clustering result.

- The weighted F-measure (*WF*) corresponds to a score resulting from the computation of F-measure for each label, and finding their average weighted by the number of true instances for each label. It takes the label imbalance into account.

These classical per-pixel classification accuracy metrics are calculated for each content type independently for each book. This shows that whether a pixel-labeling method behaves uniformly among all the books or if, conversely, it achieves a different level of performance for different books. On the other hand, computing these classification accuracy metrics for each content type helps to identify which methods have high performance for specific content types (or for one particular book), even if their overall performance is not so high. It is worth noting that computing the overall classification accuracy metrics is not a good way for evaluating the performance of an image analysis method, as it would mainly measure the performance on the majority classes due to the imbalanced headcounts between classes. The performance measures are computed for both challenges at pixel level. Both the performance measures for each book of the HBA 1.0 dataset and the overall performance are calculated. The evaluation tools used to evaluate the two challenges of this study are available to download[3].

V. Participating methods

In this section, brief descriptions of the five participating methods for both challenges are reported.

A. Barney Smith's method (M1)

Barney Smith's method (*M1*) which is proposed by Elisa H. Barney Smith from Boise State University, focuses on defining heuristics deduced from analyzing connected component, word and line characteristics.

B. Stewart and Barrett' method (M2)

Stewart and Barrett' method (*M2*) is proposed by Seth Stewart and Bill Barrett from Brigham Young University. It uses a fully convolutional network (FCN) architecture, which is based substantially on the work found in [6]. Training occurs by applying stochastic gradient descent on randomly sampled image crops with a per-class quota. Stewart and Barrett further solve the problem of class imbalance by periodically computing per-class F-measures, which are used to update the balance of each class in the minibatches, prioritizing the more difficult classes.

C. Li et al.' method (M3)

Li *et al.*' method (*M3*) which is proposed by Xiao-Hui Li, Fei Yin and Cheng-Lin Liu from the National Laboratory of Pattern Recognition (NLPR) - Institute of Automation of Chinese Academy of Sciences (CASIA) - University of Chinese Academy of Sciences, is based on a FCN model. To improve the performance of the FCN model, Li *et al.* apply the following modifications. First, to solve the problem of scale inconformity they resize the original images and ground truth maps to the same short edge of 2048 pixels. Second, for data enhancement they apply random scaling and rotation, then they randomly cut the original images and ground truth maps to small patches with size of 512×512 pixels. Third, they use weighted cross entropy as our loss function in which the weights are set as inverse class frequency in each batch. For pixels which are labeled as background in the ground truth maps but predicted as foreground by the FCN, they set their labels to background. For pixels which are labeled as foreground in the ground truth maps but predicted as background by the FCN, they use the labels of their nearest foreground neighbor as their predicted label.

D. Grüning et al.' method (M4)

Grüning *et al.*' method (*M4*), called *Argus-PixelLabeler* is proposed by Tobias Grüning, Tobias Strauß and Gundram Leifert from CITlab (University Rostock) - Planet AI GmbH. The *Argus-PixelLabeler* is based on an ARU-Net network introduced in [7]. The ARU-Net is trained to distinguish between the 6 different classes of interest. The original images are down-scaled such that they have a width of roughly 1600 pixels. Furthermore, the ground truth pixel information is dilated to increase its ratio to the entire set of pixels and to reduce errors caused by Otsu's binarization which are present in the competition data. We trained the neural network with randomly scaled versions of the training data. A simple post-processing is performed. It assigns the class of the closest labeled foreground pixel to foreground pixels, which are not labeled by the ARU-Net. No additional post-processing steps are performed.

E. Zharkov's method (M5)

Zharkov's method (*M5*) which is proposed by Andrey Zharkov from ABBYY-LLC, is based on a convolutional neural network (CNN) architecture. Zharkov trains the used CNN for semantic segmentation on entire evaluation training set and sample training set. After that he fine-tunee the weights by training with lower learning rate on each particular book for few epochs. The network architecture has 3 initial convolutions to reduce spatial image dimension by 4 then 6 more dilated convolutions with exponential increase in dilation to capture some context, then last convolution for dense predictions of class probabilities. The network is trained using mean cross-entropy loss on foreground pixels with random crops and small noise/color augmentations.

VI. Results

A brief overview of the obtained qualitative and numerical results of the five participating methods: Barney Smith's method (*M1*), Stewart and Barrett' method (*M2*), Li *et al.*'

[3] http://hba.litislab.eu/evaluation_scripts/hba_evaluation_scripts.zip

method (*M3*), Grüning *et al.*' method (*M4*) and Zharkov's method (*M5*), is presented in this section. More detailed results will be posted on the HBA Web site[1].

A. Qualitative results

Many examples of the results of the five participating methods on the HBA 1.0 dataset for the *Challenge 2* are shown in Figure 3.

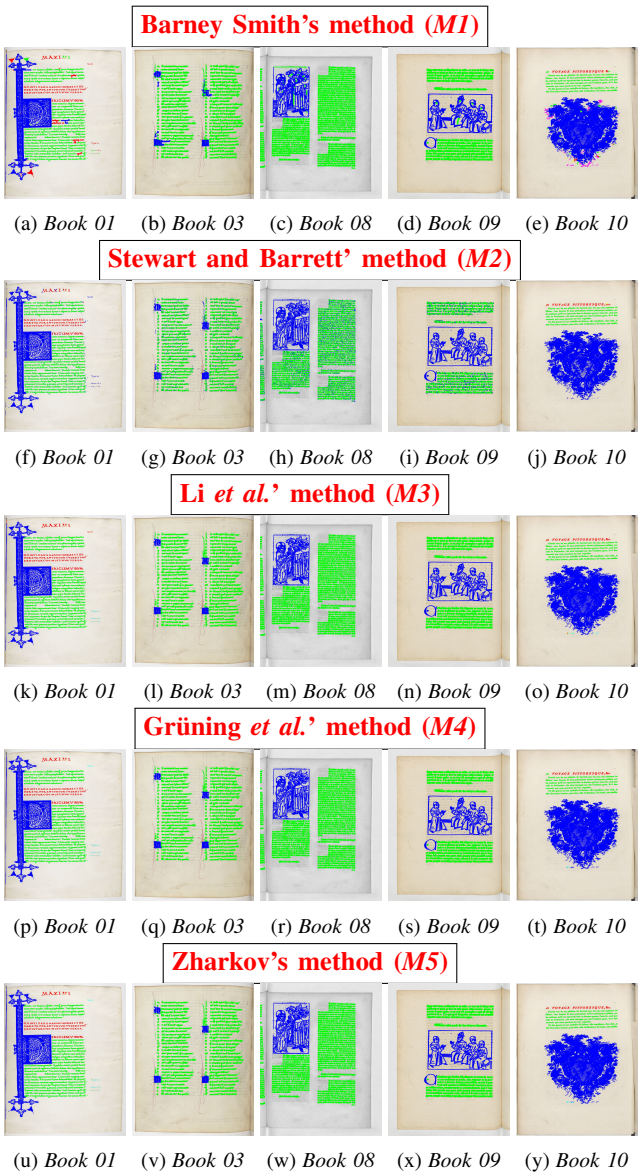

Figure 3. Examples of resulting images obtained when using Barney Smith's method (*M1*), Stewart and Barrett' method (*M2*), Li *et al.*' method (*M3*), Grüning *et al.*' method (*M4*) and Zharkov's method (*M5*) for the *Challenge 2*.

First, we note that the *M1* and *M2* methods behave differently among all the books of the HBA 1.0 dataset. For instance, we observe that the *M1* method has more difficulty separating different textual fonts (*cf.* Figure 3(a)) compared with the *M2* method (*cf.* Figure 3(f)). We also see that the textual regions (green) are more homogeneous when using the *M1* method on *Book 08* (*cf.* Figure 3(c)) and *Book 09* (*cf.* Figure 3(d)) compared with the *M2* method (*cf.* Figures 3(h) and 3(i)). Nevertheless, we observe that the graphical regions (blue) are more homogeneous when using the *M2* on *Book 08* (*cf.* Figure 3(h)) and *Book 09* (*cf.* Figure 3(i)) compared with the *M1* method (*cf.* Figures 3(c) and 3(d)). By comparing the visual results, we show that the *M3* method (*cf.* Figures 3(k) $\rightarrow$ 3(o)), the *M4* method (*cf.* Figures 3(p) $\rightarrow$ 3(t)) and the *M5* method (*cf.* Figures 3(u) $\rightarrow$ 3(y)) give the best results in terms of the homogeneity of the textual and graphical region content.

B. Quantitative results

It is necessary to assess quantitatively the results in order to have a conclusion of which method is well suited for firstly segmenting graphical regions from textual ones, and then for discriminating text in a variety of situations of different fonts and scales. The evaluation results of the five participating methods are presented in terms of the classification accuracy rate (*CA*), the F-measure (*F*) and the weighted F-measure (*WF*) on the nine books of the evaluation dataset (*Book 01*, *Book 02*, *Book 03*, *Book 05*, *Book 07*, *Book 08*, *Book 09*, *Book 10* and *Book 11*). The performance evaluation of the five participating methods in both challenges of the HBA2019 contest are presented in Table I.

First, we note that the computed accuracy values are not sufficiently congruent. Indeed, we note a slight difference in the performance of the *F* comparing the two other computed accuracy metrics (*CA* and *WF*). This is due to the non-consideration of the imbalanced headcounts between classes in the same book when the *F* metric was computed. We also observe that the best results of *WF* values are obtained by using the *M3* method on *Book 08* (99.97% and 99.98% for the *Challenge 1* and *Challenge 2*, respectively). This can be justified by the particularities of *Book 08*. Indeed, the majority of the *Book 08* pages have double-column layouts and contain single text font and graphics. Moreover, we show that the worst *WF* performances are obtained by using the *M1* method on *Book 11* (86.33%) for the *Challenge 1* and *Book 07* (61.81%) for the *Challenge 2*. This can be explained by the insufficient number of the training images of *Book 07* since it has a limited number of pages and large variability of page layout and content (*i.e.* six predefined classes in the ground truth). In addition, we note that the *M3* method achieves the best performance on *Book 01*, *Book 02*, *Book 03*, *Book 05* and *Book 07* for the *Challenge 1*, and on *Book 02*, *Book 03*, *Book 05*, *Book 07* and *Book 10* for the *Challenge 2*. On the other side, the *M4* method achieves the best performance on *Book 08*, *Book 09* and *Book 11* for the *Challenge 1*, and *Book 01*, *Book 08* and *Book 09* for the *Challenge 2*. Besides, the *M5* method

Table I
EVALUATION RESULTS OF THE PARTICIPATING METHODS.

		Challenge 1			Challenge 2		
		CA	F	WF	CA	F	WF
Book 01	M1	86.15	92.49	92.25	82.02	75.04	87.82
	M2	97.35	96.61	98.45	**97.20**	**94.74**	**98.30**
	M3	**99.96**	99.72	99.96	99.76	98.37	99.76
	M4	99.95	99.68	99.95	99.87	99.21	99.87
	M5	**99.95**	99.71	99.95	99.72	98.00	99.72
Book 02	M1	94.00	**75.81**	94.27	93.97	75.81	94.24
	M2	92.93	77.63	93.26	92.93	77.65	93.26
	M3	**99.50**	96.38	**99.51**	99.49	83.65	99.50
	M4	**99.46**	95.97	99.47	99.45	82.97	99.46
	M5	**99.37**	95.09	**99.38**	99.35	82.42	99.36
Book 03	M1	98.26	91.53	98.46	98.26	91.53	98.46
	M2	94.92	83.46	95.17	94.92	83.46	95.17
	M3	99.78	98.63	99.78	99.73	98.62	99.73
	M4	99.69	98.12	99.69	99.69	98.12	99.69
	M5	99.69	98.04	99.69	99.69	98.04	99.69
Book 05	M1	94.98	89.06	95.26	89.54	89.26	89.81
	M2	96.81	96.89	98.29	96.61	91.09	98.09
	M3	99.90	99.76	99.90	99.84	97.14	99.84
	M4	99.89	99.73	99.89	99.80	99.47	99.80
	M5	99.77	99.44	99.77	99.61	99.27	99.61
Book 07	M1	93.50	80.26	93.67	**61.70**	**43.84**	**61.82**
	M2	**98.83**	**98.39**	**99.24**	93.21	92.81	93.60
	M3	99.71	98.85	99.71	**94.11**	93.54	**94.11**
	M4	99.49	97.95	99.49	**93.94**	93.01	**93.94**
	M5	99.53	98.10	99.53	**93.28**	92.04	**93.28**
Book 08	M1	**99.01**	**98.46**	**99.17**	**99.01**	**98.46**	**99.17**
	M2	95.17	92.39	95.41	95.17	92.39	95.41
	M3	99.96	**99.93**	**99.96**	**99.96**	**99.93**	**99.96**
	M4	**99.97**	**99.97**	**99.98**	**99.97**	**99.97**	**99.98**
	M5	99.95	**99.93**	**99.96**	**99.95**	**99.93**	**99.96**
Book 09	M1	98.69	97.61	98.92	98.67	97.63	98.90
	M2	92.27	86.97	**92.38**	92.26	86.99	92.38
	M3	99.89	99.74	99.89	99.86	85.11	99.86
	M4	99.90	99.77	99.91	99.87	**82.10**	99.87
	M5	99.88	99.74	99.89	99.87	87.71	99.87
Book 10	M1	78.67	88.17	87.33	77.66	73.05	86.21
	M2	95.12	97.38	97.27	94.78	90.16	96.92
	M3	99.65	99.60	99.65	99.35	86.14	99.35
	M4	99.47	99.41	**99.47**	99.29	86.10	99.29
	M5	99.66	99.62	99.66	99.14	**81.85**	99.14
Book 11	M1	**76.22**	81.47	**86.33**	75.77	77.67	85.82
	M2	**87.29**	**47.39**	92.53	**86.80**	**41.83**	**92.01**
	M3	99.82	**93.62**	99.82	99.60	**78.89**	99.60
	M4	99.89	**95.90**	99.90	99.61	96.37	99.61
	M5	99.84	**94.04**	99.84	99.74	83.82	99.75
Overall	**M1**	**91.05**	**88.32**	**93.96**	**86.29**	**80.25**	**89.14**
	M2	**94.52**	**86.35**	**95.78**	**93.77**	**83.46**	**95.01**
	M3	**99.80**	**98.47**	**99.80**	**99.08**	**91.27**	**99.08**
	M4	**99.75**	**98.50**	**99.75**	**99.05**	**93.03**	**99.06**
	M5	**99.74**	**98.19**	**99.74**	**98.93**	**91.45**	**98.93**

M1, *M2*, *M3*, *M4* and *M5* denote the Barney Smith's method, Stewart and Barrett' method, Li *et al.*' method, Grüning *et al.*' method and Zharkov's method. *CA*, *F* and *WF* denote the classification accuracy rate, F-measure and weighted F-measure, respectively. The higher the values of the computed metrics, the better the results. The values which are quoted in **red** and **green** are considered as the **lowest** and **highest** values, respectively for each participating method. Table cells whose background are yellow note the best performance for each book with respect to the *WF* metric.

achieves the best performance on *Book 10* and *Book 11* for the *Challenge 1* and *Challenge 2*, respectively. Finally, we conclude that that the methods based on deep architectures provides better performance than the ad-hoc methods (based on different image processing techniques, such as connected component) especially when there is an insufficient number of the training images and large variability of page layout and content. Moreover, we see that the method submitted by Li *et al.* from the NLPR-CASIA team achieves the highest *WF* performance in both challenges (99.80% and 99.08% for the *Challenge 1* and *Challenge 2*, respectively) followed by Grüning *et al.* from the CITlab team (99.75% and 99.06% for the *Challenge 1* and *Challenge 2*, respectively), and Zharkov from the ABBYY-LLC team (99.74% and 98.93% for the *Challenge 1* and *Challenge 2*, respectively).

VII. CONCLUSIONS AND FUTURE PERSPECTIVES

In order to address an important lack in past benchmarking studies and contests, this work evaluates five pixel-labeling methods for HDIA. The obtained results confirm that our contest allows a consistent evaluation and a fair comparison of low-level image analysis methods for HDIA on the one hand, and that methods based on deep architectures provide better performance than the ad-hoc methods on the other hand. Moreover, the evaluation shows that the method submitted by Li *et al.* from the NLPR-CASIA team achieves the highest performance in both challenges.

An important need has been recently highlighted by researchers that consists in providing a representative region-based annotated dataset of historical document images in order to ensure the evaluation of page content classification methods at region/block level. Thus, we are currently working on developing a novel ground truthing toolkit for layout analysis of historical document images which ensures automatic generation, visualization and editing of region-based ground truths of document images based on the pixel-based ground truth of the HBA 1.0 dataset.

ACKNOWLEDGMENT

The authors would like to acknowledge the French national library (BnF) for providing access to the French digital library Gallica.

REFERENCES

[1] L. Stein and P. Lehu, *Literary research and the American realism and naturalism period: strategies and sources*. Scarecrow Press, 2009.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "ICDAR 2013 Competition on Historical Book Recognition (HBR 2013)," in *International Conference on Document Analysis and Recognition*, 2013, pp. 1459–1463.

[3] N. Serrano, F. Castro, and A. Juan, "The RODRIGO database," in *International Conference on Language Resources and Evaluation*, 2010, pp. 2709–2712.

[4] M. Mehri, P. Héroux, R. Mullot, J. P. Moreux, B. Coüasnon, and B. Barrett, "HBA 1.0: A pixel-based annotated dataset for historical book analysis," in *International Workshop on Historical Document Imaging and Processing*, 2017, pp. 107–112.

[5] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer-Verlag, 2011.

[6] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *International Conference on Document Analysis and Recognition*, 2017, pp. 99–104.

[7] T. Grüning, G. Leifert, T. Strauß, and R. Labahn, "A two-stage method for text line detection in historical documents," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1802.03345